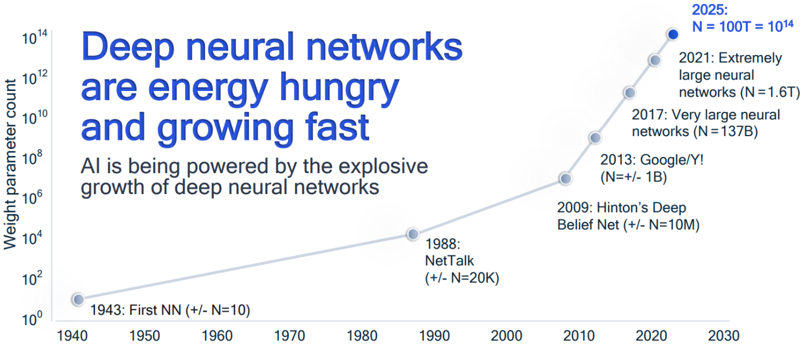
Integrating environmental impact of AI in a data center

Paul Gay

Université de Pau - Cytech

January 2024

AI impact



Deep neural networks are energy hungry and growing fast

AI is being powered by the explosive growth of deep neural networks

- Exponential growth of Deep Learning models [Fournarakis, 2021]

Planetary limits

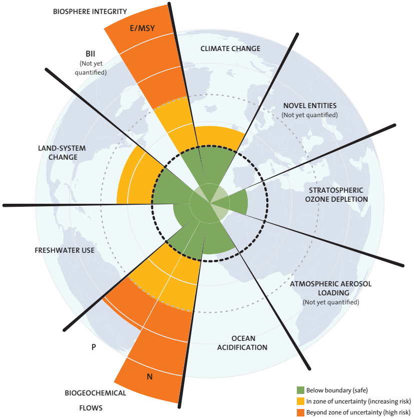


Figure 2: Credit: J. Lokrantz/Azote based on Steffen et al. 2015.

Motivation for technology

Technology can save us

- Health, environment, education, agriculture, transports, information

Usage effect

But also (ideally before), we need to wonder what is best thing to do ?

It is likely we are in a Jevons paradox

For instance, when training a deep speech model

- GPU : 47KWh over 150 hours
- CPU : 188KWh over 6000 hours
- GPU is 4 times more energy efficient than CPU!

Usage effect

But also (ideally before), we need to wonder what is best thing to do ?

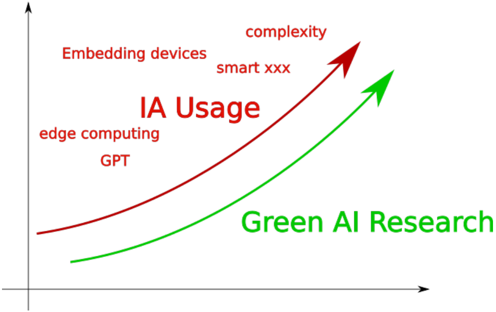
It is likely we are in a Jevons paradox

For instance, when training a deep speech model

- GPU : 47KWh over 150 hours
- CPU : 188KWh over 6000 hours
- GPU is 4 times more energy efficient than CPU!

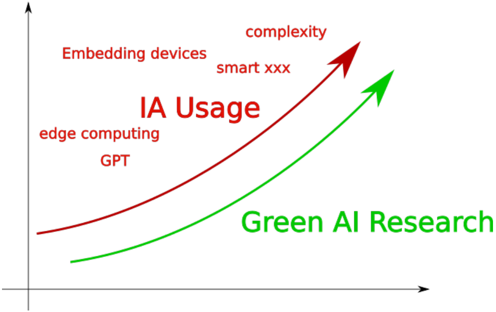
For a constant usage

Usage effect



Likely, energy efficiency increased the environmental impact of AI

Usage effect



Likely, energy efficiency increased the environmental impact of AI

Sufficiency *versus* efficiency

Measure environmental metrics

Checking the impact of the current IA

Carbon footprint estimation

- Scope 1 : Your action emits directly carbon
- Scope 2 : indirect emissions due to energy consumption : *A computer inference*
- Scope 3 : Other indirect emissions : *Large scope including manufacturing, transport, and end of life.*

+ other indirect effects (Jevons, social habits,...)

Measure environmental metrics

Checking the impact of the current IA

Carbon footprint estimation

- Scope 1 : Your action emits directly carbon
- Scope 2 : indirect emissions due to energy consumption : *A computer inference*
- Scope 3 : Other indirect emissions : *Large scope including manufacturing, transport, and end of life.*

+ other indirect effects (Jevons, social habits,...)
How to quantify all of this ?

AI impact

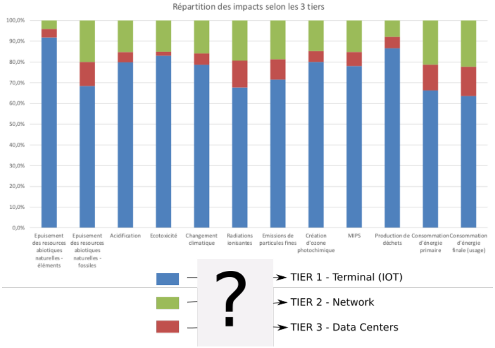


Figure 3: Multi criteria impact of French ICT. Credits Arcep/Ademe 2021

- Energy, CO2, metal usage, Water usage, ecotoxicity

AI impact

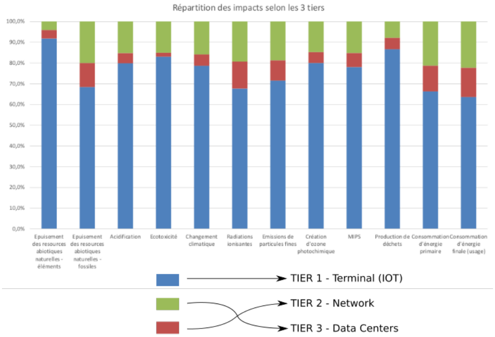
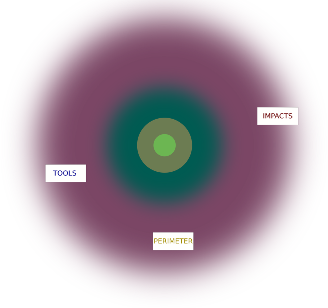


Figure 4: Multi criteria impact of French ICT. Credits Arcep/Ademe 2021

- Energy, CO2, metal usage, Water usage, ecotoxicity
- Main impact (blue) from manufacturing and IOT

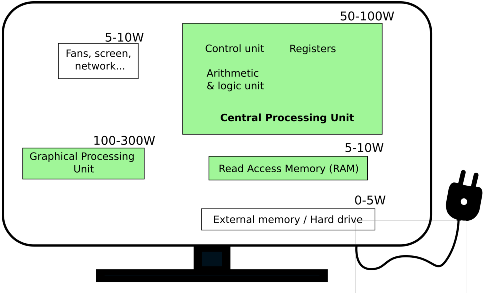
How to measure all of this ?

How to measure all of this ?



In the remaining of this presentation, let's walk through the different impacts for the different scopes, and discuss about tools to reduce them.

Computer components



5-10W
Fans, screen, network...

50-100W
Control unit
Registers
Arithmetic & logic unit
Central Processing Unit

100-300W
Graphical Processing Unit

5-10W
Read Access Memory (RAM)

0-5W
External memory / Hard drive

Easy use of internal sensors for an uncomplete consumption

```
p, q = exp.measure_yourself(period=2)
# place here the code that you want to profile
q.put(experiment.STOP_MESSAGE)
```

Scope 2 : Energy consumption at usage

Different tools based on RAPL and Nvidia-smi

- Opensource libraries for machine learning carbon footprint ([Henderson et al., 2020, Anthony et al., 2020])
- Fine grained studies on a specific Jetson hardware ([Rodrigues et al., 2018])
- French Startups : Hubblo, Dynergium

A not so trivial topic

A more rigorous approach

- Isolate the energy hungry elements
- Dependent on the built in sensor and constructor support.
- Low level (close to hardware) programming
- Energy depends on the lot of parameters

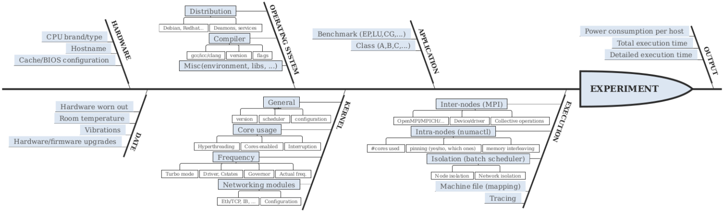


Figure 5: Picture from [Orgerie, 2020]

Inference phase

Object Detector	CNN	Vision Transformer	Text Transformer
Yolov5s	Resnet	VIT_B_16	Bert
0.61	0.27	0.94	0.07

Table 1: GPU joule consumption for one inference (check Aipowermeter doc for experimental details).

- Easy monitoring of order of magnitude
  - 10^{12} yolo inferences $\approx$ 50 km by car

Wasted joules from failed jobs

Statistics collected from the lab-ia clusters

Status	#JOBS	GPU (kWh)	CPU (kWh)	Ext. (kWh)
COMPLETED	1148	63	13	229
FAILED	134	10	8	76
CANCELLED	62	6	2	29
TIMEOUT	17	41	9	235

Table 2: Consumption per job status over 10 days and 5 machines.

Important contribution of TIMEOUT jobs

Inference phase

Many joules wasted in data scientists practices [Khan et al., 2019]

- Job crashing
- Brute force optimization to earn a few percents
- Hidden knobs and bad use of the GPUs

Inference phase

Many joules wasted in data scientists practices [Khan et al., 2019]

- Job crashing
- Brute force optimization to earn a few percents
- Hidden knobs and bad use of the GPUs

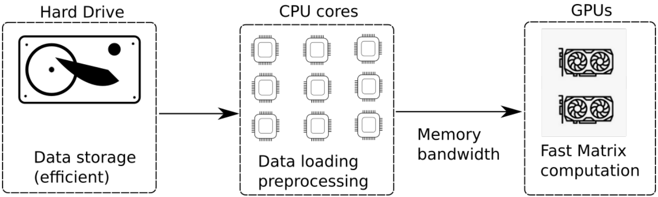
Spread good practices

- Avoid GPU bottlenecks
- Normalize your loss to avoid cuda runtime error
- Normalize your layers also
- Most of the time spent by building auxiliary code (evaluation metrics, data formatting, rect or square inference)
  - Training is slow: multiple days
  - You obtain most of the clues with small experiments and unitary tests
- Detect inefficient use of GPU

GPU monitoring

Hidden knobs and good practices for Batch Size

- GPU bottle Necks if it waits for cpu data
- Batch large enough or enough CPUs core to feed the GPU



GPU monitoring

Non linear behavior of GPU consumption

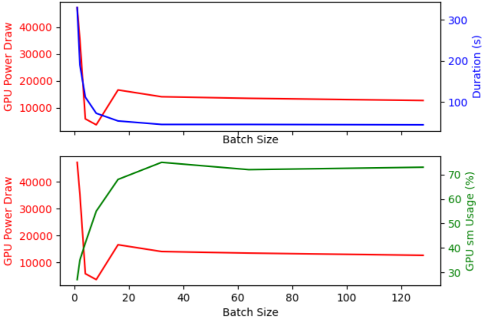


Figure 6: *Batch size* vs *Duration* vs *Power consumption* on a CNN Resnet50 with ImageNet Validation set (50000 images).

GPU usage in practice

Statistics from the lab-ia cluster ($\approx$ 9000 jobs)

- 35% of submitted jobs use GPUs
- GPUs do not seem to be used at full capacity from both memory and SM core point of view.

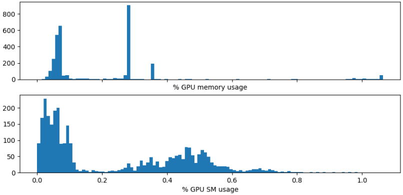


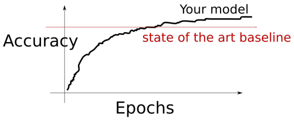
Figure 7: *% core and memory usage in GPUs*

Engagement

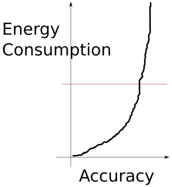
- How to create engagement ?
  - How to be legimitate ?
  - How to answer the politic question ?
  - How to avoid the "I am technical" rejection ?

Engagement

- How to create engagement ?
  - How to be legimitate ?
  - How to answer the politic question ?
  - How to avoid the "I am technical" rejection ?
- Providing metrics which matter from an environmental point of view



Common view of machine learning



Change of focus

- A playground to create dialog (about rebounce effect)
- Destroy the hype of ICT ?

Discuss consequences of IT, and how it is measured

- Energy used
- Carbon footprint
- Abiotic resources depletion
- Toxicity
- Water depletion

What about Energy Used ?

Power Usage Effectiveness (PUE)

$$PUE = \frac{\text{Total Data Center Used Energy}}{\text{IT equipment Energy}} \tag{1}$$

- Favor efficient data centers
- Favor heavily used data centers

What about carbon footprint ?

Advantages of using carbon footprint as a KPI

- **Global** change
- Hype and Popular
- More extensively studied and documented
- Identified carbon emissions [Bordage, 2019]
    - Mainly for manufacturing and usage of IOT (Arcep/Ademe 2021)

What about mineral resource depletion?

Abiotic resource depletion can be measured in *Kg Sb eq*.

- Kg Sb eq → Kilogs équivalent antimony
- which is a mineral often used to make lead
- But it is hard to interpret...

Relation with the dozens metals used in ICT is not trivial

Which metals are in ICT?

World wide production of few metals among 60 used in ICT (Estimations may vary from year and sources)

- Coppper : **28Mt**
  - 170K tons en 2017 for the French electricity network, 30K additional tons required for offshore windpower.

Which metals are in ICT?

World wide production of few metals among 60 used in ICT (Estimations may vary from year and sources)

- Coppper : **28Mt**
  - 170K tons en 2017 for the French electricity network, 30K additional tons required for offshore windpower.
- Tantale : capacitors : 1000 condensers in an iphone10. **1800 tons**

Which metals are in ICT?

World wide production of few metals among 60 used in ICT (Estimations may vary from year and sources)

- Coppper : **28Mt**
  - 170K tons en 2017 for the French electricity network, 30K additional tons required for offshore windpower.
- Tantale : capacitors : 1000 condensers in an iphone10. **1800 tons**
- Indium : screen covers **569 tons**

Which metals are in ICT?

World wide production of few metals among 60 used in ICT (Estimations may vary from year and sources)

- Coppper : **28Mt**
  - 170K tons en 2017 for the French electricity network, 30K additional tons required for offshore windpower.
- Tantale : capacitors : 1000 condensers in an iphone10. **1800 tons**
- Indium : screen covers **569 tons**
- Gallium and Germanium : power amplifiers. **320 tons and 106 tons**

Which metals are in ICT?

World wide production of few metals among 60 used in ICT (Estimations may vary from year and sources)

- Coppper : **28Mt**
    - 170K tons en 2017 for the French electricity network, 30K additional tons required for offshore windpower.
- Tantale : capacitors : 1000 condensers in an iphone10. **1800 tons**
- Indium : screen covers **569 tons**
- Gallium and Germanium : power amplifiers. **320 tons and 106 tons**

Some metals are vitamins : small quantities for more efficient products.

Source : Gaetan Lefevre La consommation croissante en matières premières du numérique : l'urgence d'une prise de conscience. 2019

Mineral Depletion ?

- The constraint comes from the market and not from the resources
- Most of the metal vitamins are subproducts of larger industries (g/tons)
  - gallium → bauxite, and indium → Zinc
- Inegality among the countries
- Competition with other usages ?
- Difficulty of recycling

What about toxicity ?

Toxicity can be measured in *CTU* (Comparative Toxic Units)

What about toxicity ?

Toxicity can be measured in *CTU* (Comparative Toxic Units)

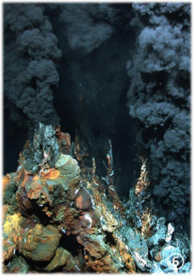
- From ILCD handbook : still immature to be recommended.

Various effects summarised in one number [Systex, 2021] :

- Mercure dams, Acid mine drainage
- Issue of the after mining period
  - Desertion and small state intervention
- Recylcing

However, these are local consequences...

Local consequences...



What about water ?

Water depletion from mining and usage in cooling systems from data centers

- Water Usage Effectiveness in Data centers: 0.25-1,8 L/kWh

What about water ?

Water depletion from mining and usage in cooling systems from data centers

- Water Usage Effectiveness in Data centers: 0.25-1,8 L/kWh
- Data Center in France : 2.7 millions m^3
- Data Center World wide : 75 millions m^3

What about water ?

Water depletion from mining and usage in cooling systems from data centers

- Water Usage Effectiveness in Data centers: 0.25-1,8 L/kWh
- Data Center in France : 2.7 millions m^3
- Data Center World wide : 75 millions m^3

Some orders of magnitude

- In France : 4,1 billions m^3 per year
- 2.3 billions m^3 for Agriculture
- 8 millions m^3 for artificial snow in Savoie in 2021
- GLENCORE : 331 millions m^3

What about water ?

Water depletion from mining and usage in cooling systems from data centers

- Water Usage Effectiveness in Data centers: 0.25-1,8 L/kWh
- Data Center in France : 2.7 millions m^3
- Data Center World wide : 75 millions m^3

Some orders of magnitude

- In France : 4,1 billions m^3 per year
- 2.3 billions m^3 for Agriculture
- 8 millions m^3 for artificial snow in Savoie in 2021
- GLENCORE : 331 millions m^3

again, these consequences are local

What about water ?

Water depletion from mining and usage in cooling systems from data centers

- Water Usage Effectiveness in Data centers: 0.25-1,8 L/kWh
- Data Center in France : 2.7 millions m^3
- Data Center World wide : 75 millions m^3

Some orders of magnitude

- In France : 4,1 billions m^3 per year
- 2.3 billions m^3 for Agriculture
- 8 millions m^3 for artificial snow in Savoie in 2021
- GLENCORE : 331 millions m^3

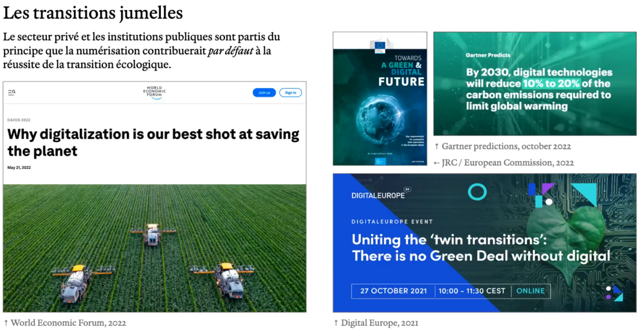
again, these consequences are local *Source : Prélevée ou consommée : comment compter (sur) l'eau ? last cheked the 29th/032023 | Commissariat général au développement durable. online article.*

What about rebound effect?

The scope of Arcep study does not include the digitalisation of the other sectors

Contradictory results from other studies

GSMA reports and the 1/10 ratio



Les transitions jumelles

Le secteur privé et les institutions publiques sont partis du principe que la numérisation contribuerait *par défaut* à la réussite de la transition écologique.

DAVOS 2022

Why digitalization is our best shot at saving the planet

May 21, 2022

↑ World Economic Forum, 2022

TOWARDS A GREEN & DIGITAL FUTURE

Gartner Predicts

By 2030, digital technologies will reduce 10% to 20% of the carbon emissions required to limit global warming

↑ Gartner predictions, october 2022

← JRC / European Commission, 2022

DIGITALEUROPE

DIGITALEUROPE EVENT

Uniting the 'twin transitions': There is no Green Deal without digital

27 OCTOBER 2021 | 10:00 - 11:30 CEST | ONLINE

↑ Digital Europe, 2021

Figure 8: Slide from Gauthier Roussilhe, *La numérisation aide-t'elle la transition écologique* https://labos1point5.org/les-seminaires

- smart building (-40%), smart agriculture (-65%), remote medicine, airbnb,...

Problems of estimating rebounce effect [Rasoldier et al., 2022]

- World wide prospectives unreliable
- Comparison with a worst case scenario
- Only one environmental factor is considered
- End of cycle is difficult to take into account

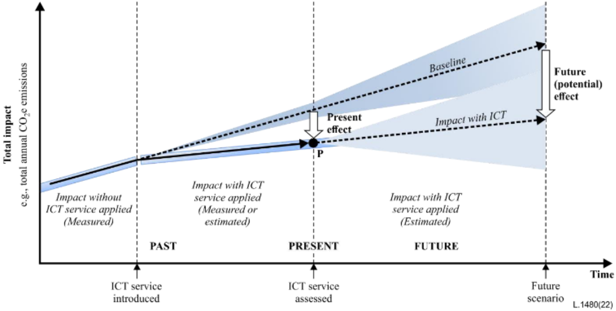
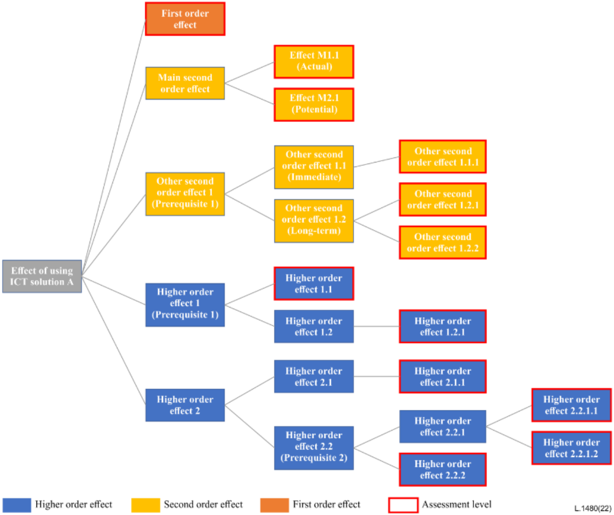

Total impact
e.g., total annual CO$_2$e emissions
Baseline
Future (potential) effect
Present effect
Impact with ICT
P
Impact without ICT service applied (Measured)
Impact with ICT service applied (Measured or estimated)
Impact with ICT service applied (Estimated)
PAST
PRESENT
FUTURE
Time
ICT service introduced
ICT service assessed
Future scenario
L.1480(22)

Figure 9: Recommendation ITU-T L.1480. 2022

Impact assessment of AI project

Environmental assessment of projects involving AI methods [Lefèvre et al., 2023]

- Estimate the impact of your research
- including:
  - Devices for initial training
  - Devices while in production
  - Life cycle assessment
  - Resilience, quality of service
  - Prospective exercice : imagine the future with consequence tree



Effect of using ICT solution A

First order effect

Main second order effect

Effect M1.1 (Actual)

Effect M2.1 (Potential)

Other second order effect 1 (Prerequisite 1)

Other second order effect 1.1 (Immediate)

Other second order effect 1.1.1

Other second order effect 1.2 (Long-term)

Other second order effect 1.2.1

Other second order effect 1.2.2

Higher order effect 1 (Prerequisite 1)

Higher order effect 1.1

Higher order effect 1.2

Higher order effect 1.2.1

Higher order effect 2

Higher order effect 2.1

Higher order effect 2.1.1

Higher order effect 2.2 (Prerequisite 2)

Higher order effect 2.2.1

Higher order effect 2.2.1.1

Higher order effect 2.2.1.2

Higher order effect 2.2.2

🟦 Higher order effect 🟨 Second order effect 🟧 First order effect □ Assessment level

L.1480(22)

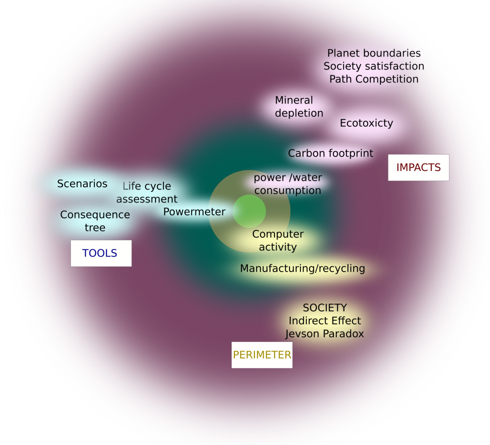
Figure 10: Recommendation ITU-T L.1480. 2022

Méthodologie impact carbone de l'Ademe

EMPREINTE PROJET : ÉVALUER L'EMPREINTE ENVIRONNEMENTALE D'UN PROJET (ADEME, 2023)

- Analyse en cycle de vie
  - Définir un périmètre
  - Identifier les principaux impacts et critères
  - Construire un arbre de conséquences
  - Quantifier ce qui est possible

The impact of your work



- Accurate measurement at lower scope

Another example: what is the impact of your research

You Only Look Once *versus* You only live once

- 10^{12} yolo inferences $\approx$ 50 km by car
- Let's assume a researcher drive to the lab 400 days over 2 years.
- If he invents an algorithm to divide yolo consumption by 2 AND industrialise his discovery to process 4 millions hours of video (1 day of french youtube).
- He makes up for his car driving ...
- ... if the saved energy is not used for something else.

References I

[Anthony et al., 2020] Anthony, L., Kanding, B., and Selvan, R. (2020).
Carbontracker: Tracking and predicting the carbon footprint of training deep learning models.
page arXiv preprint https://arxiv.org/abs/2007.03051.

[Bordage, 2019] Bordage, F. (2019).
Empreinte environnementale du numérique mondial.
Paris, greenit. fr, page 9.

[Fournarakis, 2021] Fournarakis, M. (2021).
A practical guide to neural network quantization.

[Henderson et al., 2020] Henderson, P., Hu, J.-R., Romoff, J., Brunskill, E., Jurafsky, D., and Pineau, J. (2020).
Towards the systematic reporting of the energy and carbon footprints of machine learning.
ArXiv, abs/2002.05651.

[Khan et al., 2019] Khan, K., Scepanovic, S., Niemi, T., Nurminen, J., Von Alfthan, S., and Lehto, O. (2019).
Analyzing the power consumption behavior of a large scale data center.
SICS Software-Intensive Cyber-Physical Systems, 34:61–70.

[Lefèvre et al., 2023] Lefèvre, L., Ligozat, A.-L., Trystram, D., Bouveret, S., Bugeau, A., Combaz, J., Frenoux, E., Guennebaud, G., Lefèvre, J., Nicolaï, J.-P., and Dassas, K. (2023).
Environmental assessment of projects involving AI methods.
working paper or preprint.

References II

[Orgerie, 2020] Orgerie, A.-C. (2020).
From Understanding to Greening the Energy Consumption of Distributed Systems.
PhD thesis, Ecole Normale Supérieure de Rennes.

[Rasoldier et al., 2022] Rasoldier, A., Combaz, J., Girault, A., Marquet, K., and Quinton, S. (2022).
How realistic are claims about the benefits of using digital technologies for ghg emissions mitigation?

[Rodrigues et al., 2018] Rodrigues, C. F., Riley, G., and Luján, M. (2018).
Synergy: An energy measurement and prediction framework for convolutional neural networks on jetson tx1.
In *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA)*, pages 375–382. The Steering Committee of The World Congress in Computer Science, Computer ....

[Systex, 2021] Systex (2021).
Controverses minières pour en finir avec certaines contrevérités sur la mine et les filières minérales.